Assignment 3 STAT 315-463: Multivariable Statistical Methods and Applications

Dataset manipulation:

```
contraception = read.csv("Contraception315.csv", header = TRUE)

# Convert categorical variables to factors
contraception$district <- as.factor(contraception$district)
contraception$urban <- as.factor(contraception$urban)
contraception = contraception[,-1]
```

Question 1

```
poisson <- glm(livch ~ ., data = contraception, family = "poisson")
summary(poisson)
```

```
##
## Call:
## glm(formula = livch ~ ., family = "poisson", data = contraception)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6302  -1.2605  -0.2517   0.5293   3.5116
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.7054865  0.0742477   9.502  < 2e-16 ***
## district6   -0.1072500  0.1042029  -1.029  0.30337
## district14  -0.2638122  0.0967032  -2.728  0.00637 **
## district25   0.0006459  0.1038326   0.006  0.99504
## district46  -0.0930164  0.0971626  -0.957  0.33840
## useY         0.3474549  0.0676195   5.138 2.77e-07 ***
## age          0.0655158  0.0036114  18.141  < 2e-16 ***
## urbanY      -0.1782521  0.0787776  -2.263  0.02365 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 991.36  on 452  degrees of freedom
## Residual deviance: 618.13  on 445  degrees of freedom
## AIC: 1592.7
##
## Number of Fisher Scoring iterations: 5
```

Question 2

```r
library(MASS)
poisson_backward <- stepAIC(poisson, direction = "backward")
```

```
## Start:  AIC=1592.67
## livch ~ district + use + age + urban
##
##            Df Deviance    AIC
## <none>          618.13 1592.7
## - district  4   627.19 1593.7
## - urban     1   623.30 1595.8
## - use       1   644.47 1617.0
## - age       1   946.05 1918.6
```

```r
summary(poisson_backward)
```

```
##
## Call:
## glm(formula = livch ~ district + use + age + urban, family = "poisson",
##     data = contraception)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6302  -1.2605  -0.2517   0.5293   3.5116
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.7054865  0.0742477   9.502  < 2e-16 ***
## district6   -0.1072500  0.1042029  -1.029  0.30337
## district14  -0.2638122  0.0967032  -2.728  0.00637 **
## district25   0.0006459  0.1038326   0.006  0.99504
## district46  -0.0930164  0.0971626  -0.957  0.33840
## useY         0.3474549  0.0676195   5.138 2.77e-07 ***
## age          0.0655158  0.0036114  18.141  < 2e-16 ***
## urbanY      -0.1782521  0.0787776  -2.263  0.02365 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 991.36  on 452  degrees of freedom
## Residual deviance: 618.13  on 445  degrees of freedom
## AIC: 1592.7
##
## Number of Fisher Scoring iterations: 5
```

After applying backwards selection we can derive the equation:

log(livch) = 0.7054865 - 0.1072500 * district6 - 0.2638122 * district14 + 0.0006459 * district25 - 0.0930164 * district46 + 0.3474549 * useY + 0.065893 * age - 0.1782521 * urbanY

Question 3

```r
confint(poisson_backward)
```

```
## Waiting for profiling to be done...

##                    2.5 %      97.5 %
## (Intercept)  0.55767507  0.84876661
## district6   -0.31323548  0.09553372
## district14  -0.45390751 -0.07465288
## district25  -0.20445551  0.20282526
## district46  -0.28407465  0.09698992
## useY         0.21490627  0.48002839
## age          0.05844598  0.07260488
## urbanY      -0.33332245 -0.02447621
```

Use: for every 1 unit increase in useY, Livch increases by approximately exp(0.3474549) = 1.41546 or 41.546%, CI = (0.21490627 0.48002839)

Age: for every 1 unit increase in age, Livch increases by approximately exp(0.0655158) = 1.06771 or 6.771%, CI = (0.05844598 0.07260488)

Urban: for every 1 unit increase in urbanY, Livch decreases by approximately exp(-0.1782521) = 0.8367315 or 16.326%, CI = (-0.33332245 -0.02447621)

Question 4

Over dispersion occurs when the observed variance is greater than the variance predicted by the model. This means that our model may underestimate, this often occurs when there are confounding variables or factors that effect the outcome.

```r
phi=sum((resid(poisson_backward,type="pearson")^2)/(poisson_backward$df.residual));phi
```

```
## [1] 1.311184
```

To account for overdispersion in our model, we can fit a quasi-Poisson:

```r
quasipoisson <- glm(livch ~ ., data = contraception, family=quasipoisson())
summary(quasipoisson)
```

```
##
## Call:
## glm(formula = livch ~ ., family = quasipoisson(), data = contraception)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6302  -1.2605  -0.2517   0.5293   3.5116
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  0.7054865  0.0850190   8.298 1.28e-15 ***
## district6   -0.1072500  0.1193200  -0.899   0.3692
## district14  -0.2638122  0.1107322  -2.382   0.0176 *
## district25   0.0006459  0.1188960   0.005   0.9957
## district46  -0.0930164  0.1112583  -0.836   0.4036
## useY         0.3474549  0.0774293   4.487 9.19e-06 ***
## age          0.0655158  0.0041353  15.843  < 2e-16 ***
## urbanY      -0.1782521  0.0902062  -1.976   0.0488 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.311193)
##
##     Null deviance: 991.36  on 452  degrees of freedom
## Residual deviance: 618.13  on 445  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

Looking at phi, our over dispersion parameter is above 1, which suggests there is over dispersion in our dataset. For no over dispersion to occur this parameter must be 1. To improve our accuracy, we can create a quasi-Poisson model, running the code above gives us such model. The quasi-Poisson model assumes that the variance is proportional to the mean, this allows for some more variation in the response variable, and reducing the risk of type I errors in hypothesis testing.

Question 5

```r
contraception$child <- ifelse(contraception$livch > 0, 1, 0)
```

Question 6

```r
logistic <- glm(child ~ district + use + age + urban, data = contraception, family = "binomial")
summary(logistic)
```

```
##
## Call:
## glm(formula = child ~ district + use + age + urban, family = "binomial",
##     data = contraception)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8275  -0.5029   0.2166   0.5870   2.0520
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.41356    0.40464   5.965 2.45e-09 ***
## district6   -0.58600    0.49453  -1.185 0.236033
## district14  -0.49151    0.39921  -1.231 0.218252
## district25  -0.22601    0.47423  -0.477 0.633661
```

```
## district46  -0.85674    0.47970  -1.786 0.074100 .
## useY         1.18373    0.30810   3.842 0.000122 ***
## age          0.26015    0.02926   8.892  < 2e-16 ***
## urbanY      -0.89039    0.36311  -2.452 0.014202 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 515.43  on 452  degrees of freedom
## Residual deviance: 330.90  on 445  degrees of freedom
## AIC: 346.9
##
## Number of Fisher Scoring iterations: 6
```

```
options(scipen=999)
```

Question 7

```
logistic_backward <- stepAIC(logistic, direction = "backward")
```

```
## Start:  AIC=346.9
## child ~ district + use + age + urban
##
##            Df Deviance    AIC
## - district  4   334.93 342.93
## <none>          330.90 346.90
## - urban     1   337.06 351.06
## - use       1   346.67 360.67
## - age       1   487.79 501.79
##
## Step:  AIC=342.93
## child ~ use + age + urban
##
##         Df Deviance    AIC
## <none>       334.93 342.93
## - urban  1   340.99 346.99
## - use    1   347.89 353.89
## - age    1   495.00 501.00
```

```
summary(logistic_backward)
```

```
##
## Call:
## glm(formula = child ~ use + age + urban, family = "binomial",
##     data = contraception)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7498  -0.5039   0.2276   0.5950   1.9494
```

```
##
## Coefficients:
##             Estimate Std. Error z value             Pr(>|z|)
## (Intercept)  1.97780    0.26611   7.432    0.000000000000107 ***
## useY         1.02452    0.29181   3.511             0.000446 ***
## age          0.26172    0.02916   8.976 < 0.0000000000000002 ***
## urbanY      -0.69073    0.28373  -2.434             0.014916 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 515.43  on 452  degrees of freedom
## Residual deviance: 334.93  on 449  degrees of freedom
## AIC: 342.93
##
## Number of Fisher Scoring iterations: 6
```

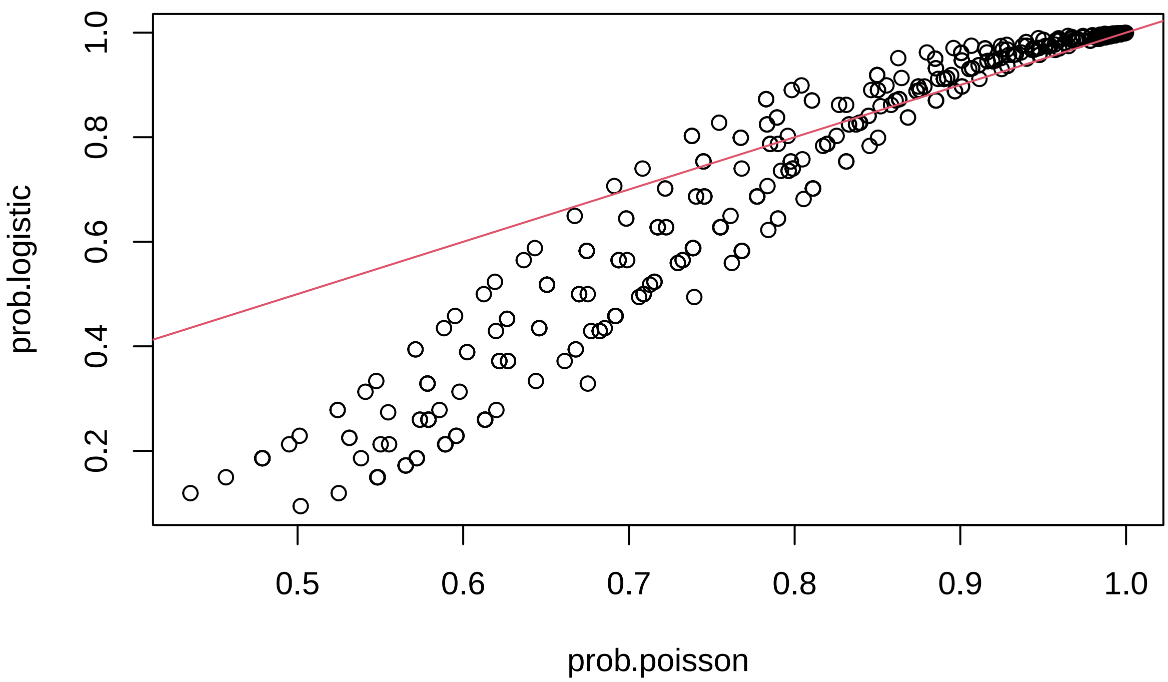
Applying backwards selection we are left with just use age and urban predictors.

log(child) = 1.97780 + 1.02452 * use + 0.26172 * age - 0.69073

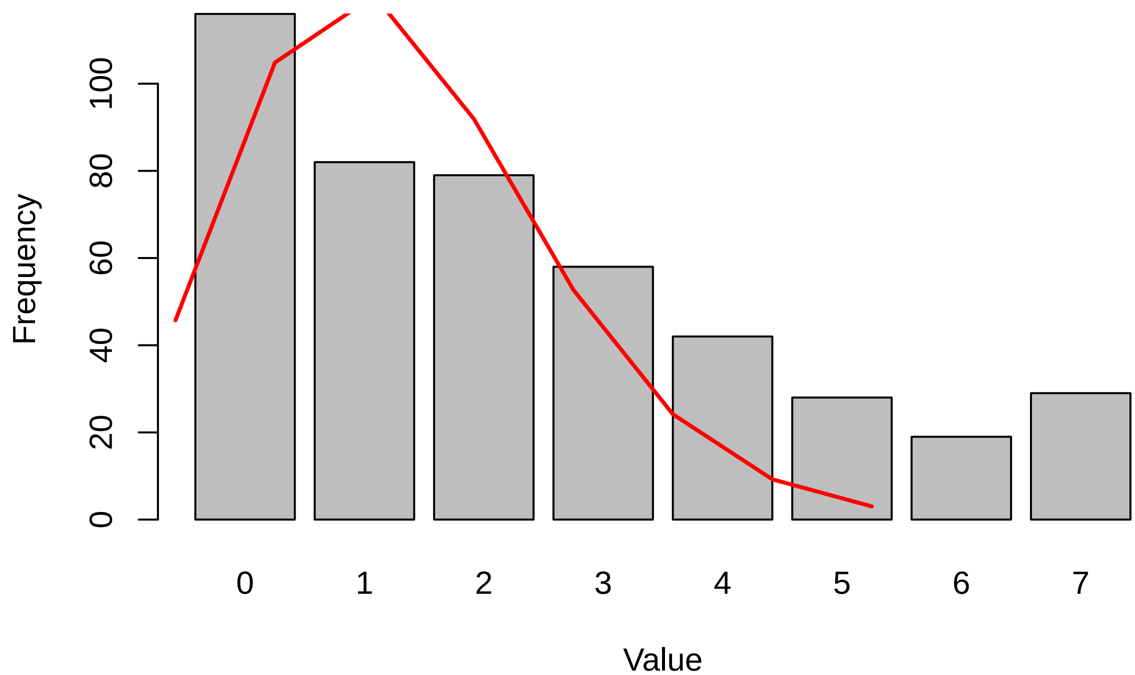
This model does not include district as a predictor after backwards selection, whereas the poisson model does.

Question 8

```
lambda <- (predict(poisson_backward, type = 'response'))
prob.poisson <- 1 - exp(-lambda)
prob.logistic <- predict(logistic_backward,type='response')
plot(prob.poisson,prob.logistic)
abline(0,1,col=2)
```




```r
cont_df <- as.data.frame(table(contraception$livch))

livch_mean = mean(contraception$livch)
livch_var = var(contraception$livch)

x_axis = seq(min(contraception$livch), max(contraception$livch), by = 1)

prob.poisson2 = dpois(x_axis, lambda = livch_mean)

barplot(cont_df$Freq, names.arg = cont_df$Var1, ylab = "Frequency", xlab = "Value")

lines(x_axis, prob.poisson2 * sum(cont_df$Freq), col = "red", lwd = 2)
```



We can see that our model does not follow a poisson distribution. This is because there are too many values of 0 in our data. Because there are so many 0's, in the lower half of the graph the poisson probabilities are higher than the logisitic, this is why the logistic/poisson graph does not follow the red line in the beginning. Despite such proabability difference in the lower half, by ~0.8 the graphs begin to coincide. The bar graph above displays the significance of the '0' inflation.